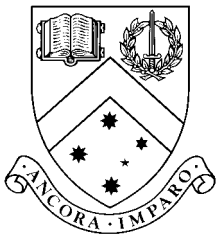
ISSN 1440-771X
ISBN 0 7326 1059 1

MONASH UNIVERSITY



AUSTRALIA

The Predictive Approach to Teaching Statistics

Alan McLean

Working Paper 4/99
March 1999

DEPARTMENT OF ECONOMETRICS AND BUSINESS STATISTICS

THE PREDICTIVE APPROACH TO TEACHING STATISTICS

Alan McLean, Monash University[1]

Abstract:

Statistics is commonly taught as a set of techniques to aid in decision making, by extracting information from data. It is argued here that the underlying purpose, often implicit rather than explicit, of every statistical analysis is to establish a set of probability models which can be used to predict values of one or more variables. Such a model constitutes 'information' only in the sense, and to the extent, that it provides predictions of sufficient quality to be useful for decision making. The quality of the decision making is determined by the quality of the predictions, and hence by that of the models used.

Using natural criteria, the ‘best predictions’ for nominal and numeric variables are respectively the mode and mean. For a nominal variable, the quality of a prediction is measured by the probability of error; for a numeric variable, it is specified using a prediction interval.

Presenting statistical analysis in this way provides students with a clearer understanding of what a statistical analysis is, and its role in decision making[2].

Keywords: Statistics, teaching, prediction, probability model, prediction interval.

[1] A presentation based on an early version of this paper was made at ICOTS-5, Singapore, June 1998, and is available in the Proceedings of that conference.

[2] It can be argued that all analysis involves creating a model of the real world. A statistical analysis is one which involves mathematical, probability models.

1. Introduction

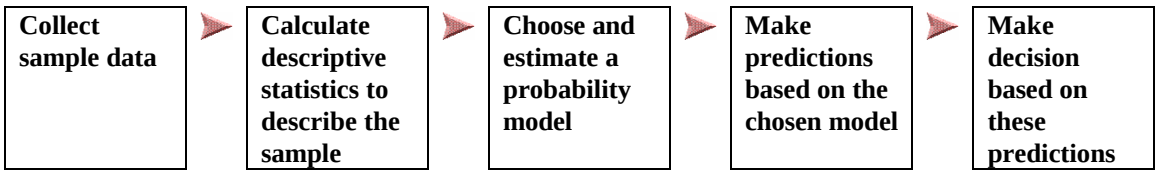
There seems to be little discussion in the statistics education literature as to what statistics ‘is’, and its influence on the way the subject could be taught. This paper is the result of many years’ teaching and thinking about the subject, and the influence of many now forgotten conversations with colleagues and students.

The typical introductory text in business statistics claims, in one way or another, that the use of statistics is to aid in decision making in conditions of uncertainty. This is certainly true, but few texts do much, in any general way, to show how statistical analysis does aid in decision making. Yet there is an underlying structure to the use of statistics which provides a unifying theme for the subject, but which is rarely made apparent to students. This underlying theme can be called the Predictive Voice.

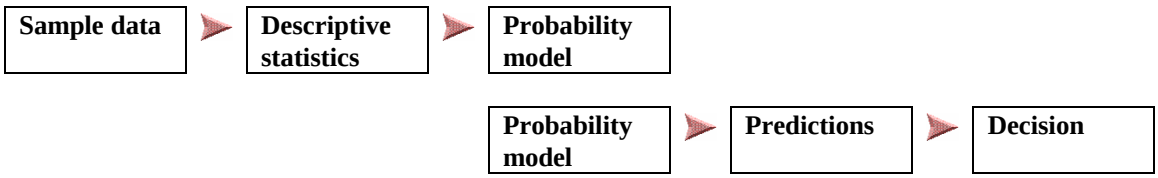
Decisions are made on the basis of what is considered likely to happen in the future if particular choices are made. They are thus based on the use of probability models. (If there is no uncertainty about the outcome of a choice, the probability distribution is degenerate.) In everyday life, these models are intuitive, based very much on personal experience and, often, personal prejudice. In more formal decision making, they may be more objective. The use of statistical analysis is to generate or validate these models, and to provide estimates of how good forecasts based on these models are likely to be. Decisions are made on the basis of the predictions.

The key word is 'model'. There is no such thing in the 'real world' as probability; but statistics is not about finding 'truth' but about finding useful ways of describing 'reality'.

The process of decision making based on statistics can be expressed diagrammatically as:



This process comprises two stages, which we consider in turn:



In short, every practical statistical analysis (including an unconscious one) is directed toward establishing a probability model of an appropriate level of complexity, which can then be used to make 'predictions' in some sense, on the basis of which decisions can be made. This simple observation provides an overall theme for the study of statistics. This paper draws out some of its implications. For example, the goal of making predictions determines the choice of what summary statistics we use.

2. The role of prediction

2.1 Basic vocabulary

Many of the problems of students originate from an inadequate knowledge of the basic vocabulary, reflecting a lack of understanding of the concepts encapsulated in the words. The terms used vary, but the important ones are briefly as follows.

A set of data may be collected as a **snapshot**, describing a set of entities at a point in time, or it may be collected as one or more **time series**, describing a single entity over time, or as a **longitudinal study** (or **panel data**), describing a set of entities over time. The set of entities forms a **population**, or a subset of a population, called a **sample**. The data comprise **values** of one or more **variables**, each of which describes some characteristic of the entities. (Most of the real problems with the use of statistics arise because of the distinction between such characteristics and the variables used to measure them.) In the case of time series data, the variation is over time; for snapshot data it is over the members of the population; for panel data it is over both time and entities. Finally, the variables are distinguished according to their nature. The most useful classification is to identify a variable as **nominal** (the values are simply labels), **ordinal** (the values have some order) or **numeric** (the values have both order and scale).

For each variable the ‘next’ value is uncertain. In the case of time series, this is because the next value is to be measured in the future. For snapshot data, it is because the next value is to be measured on an entity which is yet to be selected - so although the value is in a sense already known, its identification is in the future. It is clear that the means by which the entity is to be selected is of crucial importance in the nature of the variation. In statistics it is of course usual to assume that the selection is random.

For both time series and snapshot data, the variable is a **random variable**. For the time series case, this simply means that the next value is unpredictable; for the snapshot case, it means that the **selection of the entity** is unpredictable; more strictly, that the selection is random. Note that a random variable need not be numeric.

Probability is concerned with an action whose outcome is uncertain. If the set of outcomes is clearly identified, the action is called an **experiment**. In everyday life probability concepts (if not theory!) are used without all outcomes being clearly formulated - 'If I run across the road now I am likely to be knocked over!'

Observing the outcome of an action is synonymous with measuring a random variable. The variable may be nominal or numeric, and the action may be part of a series in which the set of possible outcomes is more or less unchanging, or it may be unique. A simple example of a unique nominal variable is the result of the next race, with values ‘horse A’, ‘horse B’, etc. A nominal time series example is the result of the next toss of a coin, with values ‘heads’ and ‘tails’.

Probability theory and descriptive statistics both are concerned with random variables. In probability, the concern is to predict the future results of an action, and hence to make decisions. In descriptive statistics the concern is to analyse the results of that action when it was performed in the past. For a variable whose values form a time series, this observation is obvious.

For snapshot data, where the variation is over the members of the population, and the uncertainty arises from the selection process, it is perhaps less obvious. From the probability viewpoint, the interest is to make a statement about the likely result of future random selections. From a descriptive statistics viewpoint, the interest is in describing a set

of people previously measured – either the whole population, or a sample from that population.

The fundamental question, and the one that drives statistical analysis, is: what is the purpose of this description of past values?

2.2 Prediction

In the case of the time series example, the answer is clear: the description of past values is used to establish a probability model so that future values can be predicted. One may argue that the purpose is to ‘understand’ the behaviour of the time series, but the only reason for doing this is to predict its future values.

For snapshot data, the answer is not so familiar. To many, the description – of the population - is itself the goal, using inference techniques if the data are measured on a sample. In a classroom this may be fine, but in a practical situation it is inadequate. How is the description to be used? Clearly, to make predictions about members of the population.

For example: An airline spaces its economy seating so that people less than 1.8 metres tall are comfortable. What is the probability that a randomly selected traveller will be comfortable? This question can be asked as: What proportion of travellers are expected to be comfortable? and it is true that most people find this a more comprehensible way of expressing it. But it cannot be asked as: What proportion of travellers **will** be comfortable? The question is a probability question, because of the uncertainty about which individuals are flying.

The description of the population, then, is used to establish a probability model which can be used to ‘predict’ results such as the proportion of travellers who will be uncomfortable.

The word 'predict' is used here in the general sense of a probability statement as illustrated above. Many authors use it for the particular technical application of projecting a time series into the future. The present more general usage encompasses time series forecasting as well as deterministic prediction. What it does not cover is the reading of crystal balls, tea leaves or entrails. It does include unconscious prediction and it is important that students realise this.

2.3 Probability Models

It is most important to convey to students that there is in the real world no such thing as 'probability' or a 'probability distribution'. A probability distribution is a **model** of the real world, on the basis of which a prediction - in the sense stated above - can be made.

A probability distribution is a **model for the future**: it lists the outcomes which are accepted as possible, and these probabilities measure, according to the model, how likely each outcome is to occur. To say that the probability of a coin coming up heads is 0.5, is not to express some absolute truth, but to describe a model which experience has shown to be useful.

The concept of a probability model becomes more explicit when the standard distributions are introduced. Students learn the conditions, for example, under which a binomial model is (likely to be) applicable. Again, it is important that students learn that the binomial is only a model which sometimes describes a real world situation quite well.

For the normal model, teachers typically rely on 'experience shows that this type of variable has (approximately) a normal distribution…..'. Students, when they ask about behaviour at infinity, may force the answer that normality ‘is a good model near the centre of the distribution.'

One can argue that there is some real but unknown probability distribution for any experiment, but this is meaningless, since even if it exists it is unknown. Rather, there is a complex deterministic process underlying the action, which is being replaced by a manageable probabilistic model. In tossing a coin, if one knew everything about how the coin was held, at what angle, how hard it was flipped and in what direction the thumb moved, the distribution of density within the coin - the result of the toss could be predicted. For practical purposes, this computation is impossible. It is replaced by the standard model: the coin is equally likely to come down heads or tails.[3]

2.4 Sources of models

The typical textbook introduction to probability identifies a spectrum of ways in which probabilities are arrived at, with ‘subjective probability’ at one end, and ‘objective

[3] This is described very well in DeGroot.

probability' based on long run proportions and exemplified in games of chance at the other. The middle part of the spectrum is occupied by 'frequentist probability', based on short run proportions. Rarely is it pointed out, first, that there is **always** some element of subjectivity, and second, that in using short run proportions one is **inferring** on the basis of sample data.

With subjective probabilities it is clear that a model is being formulated, and whether or not it is valid can be discussed. With a game of chance, it has to be emphasised that the game must be **fair**; that is, the die or coin is balanced, the cards are well shuffled. Then it can reasonably be assumed that all outcomes are equally likely – and we have our model.

Despite the rise of the casino culture, the importance of this equal probability model is not with games of chance, but in that it underlies random sampling. If a variable is measured on a population, the 'probability distribution of the variable' refers to the probability of each value **when a member of the population is selected randomly**.

If a variable has been measured for all members of a population, the probability distribution is known, and this can be used, perhaps with some simplification through grouping of data, as an 'empirical model'. Alternatively, the empirical model can be approximated by some standard model. Conceptually the empirical model is simpler, but is likely to be computationally more intense.

Whatever the source of the model, there is always a subjective element in its choice. And whether or not the choice was good is eventually determined by whether or not it works!

3. Probability and Prediction

3.1 The best prediction for a nominal variable[4]

Recognising that a probability distribution is used to predict what will happen when the experiment is carried out, what is the 'best' prediction? This of course depends on the criterion used, which in turn depends on the type of variable.

For a nominal variable, it is reasonable to **minimise the probability of error**, and define the best prediction as the outcome which is most likely to eventuate; that is, the **mode**.

[4] Foddy (1988); Freeman (1965) introduce the idea of an avereage as a 'best guess'

This use of the mode has nothing to do with 'centrality' – in any case this concept is meaningless with a nominal variable. Note that there may not be a single 'best forecast'. For a nominal variable, it is necessary to know the probability of each outcome in order to determine the mode. For a variable defined on a population, under random sampling, it is necessary to know the proportion of the population for each value of the variable.

How good is the best forecast? With a nominal variable, the quality of the forecast is automatically specified by giving the probability of it being correct. This is for most people a meaningful way of expressing the result. ‘I predict that it will rain tomorrow. The probability of rain is 0.8.’

One needs to be careful here. This is a conditional prediction, the best forecast in terms of the model, which can be expressed as

Outcome	**Probability**
Rain	0.8
Not rain	0.2

Both the prediction and its quality, in this sense, are based on the chosen model. The prediction is more carefully expressed as: ‘According to the model I have used it will rain tomorrow with probability 0.8.’ This model may or may not be a good model. Even taking into account the past performance of the model with a tree diagram approach gives a conditional prediction!

3.2 The best prediction for a numeric variable

With a numeric variable, if the number of different outcomes is small, it is again reasonable to **minimise the probability of error**, so that the best prediction is the **mode**.

However, the numeric scale gives the option of using the concept of **forecasting error**, with the best forecast as that which in some way **minimises the likely error**. If the variable has many possible values, this option must be taken.

The almost invariable choice is to minimise both the **absolute expected error** and the **expected squared error**; this is achieved by using the **mean** as the prediction. If the mean is used as the forecast, the absolute expected error is zero, ensuring that the prediction has

no bias built in, and the expected squared error is just the variance. In comparing forecasts across variables, models or populations, if in each case the mean is used, the forecasts will have zero expected error. The best forecast will then be the one with the smallest variance.

The reason for the choice of squared errors is usually explained in terms of 'getting rid of the negative signs'; that this is better than using absolute errors because it is 'mathematically more amenable'. These reasons are plausible, but the real reason is that mathematics is largely based on models in which variables are assumed to be mutually orthogonal; that is, on Pythagoras. There is no reason, in principle, why some other criterion, or **loss function**, cannot used; for example, the expected absolute error.

3.3 Prediction intervals

For a numeric variable, assuming the mean is used as prediction, the absolute expected error is zero, so forecast quality is measured by the expected squared error. This is generally not meaningful in practical terms. This is in contrast to the nominal case, where the forecast quality is measured by the probability of error. A more practically meaningful measure does exist: specify a **prediction interval** – an interval in which the result will lie with specified probability.

Prediction intervals can be calculated for any distribution, but this is rarely done in the textbooks. A similar type of calculation provides exercises for the normal distribution, but the concept of a prediction interval is not developed. For a normally distributed variable, a (symmetric) prediction interval has limits

$$m \pm z_{\alpha/2} \times s$$

For a time series, the prediction interval is immediately understood. If daily demand for a product is normally distributed with mean 200 and standard deviation 20, the predicted demand for tomorrow is 200. A 95% prediction interval on this forecast has limits

$$200 \pm 1.960 \times 20 = 200 \pm 3.9 = 196.1 \text{ to } 183.9$$

If the model is a reasonable description of reality, that is if it is 'valid', tomorrow's demand will be within this interval with probability 0.95. Conversely, there is a 5% chance that the error in prediction will be greater than 3.9.

The usefulness of the prediction interval in this case is quite clear. Less clear is the case of a variable measured on a population. If a population of heights are normally distributed with mean 170 cm and standard deviation 10 cm, predict the height of a randomly chosen member of the population as 170 cm. A 95% prediction interval on this forecast has limits

$$170 \pm 1.960 \times 10 = 170 \pm 19.6 = 150.5 \text{ to } 189.6$$

If the model is valid, there is a 95% chance that a randomly chosen person's height will be in this range. It also means that 95% of the population have heights in this range.

It must be emphasised that, as with all probability, the prediction interval is only meaningful to the extent that the model on which it is based accurately reflects reality.

Variability and prediction error are intimately related. The latter is due to the former, appreciated particularly in the time series case. In the snapshot case, the prediction error gives students another way of understanding variation and the standard deviation.

4. Descriptive Statistics and Probability

It was pointed out in Section 1 that the decision process consists of two stages, in the first of which an appropriate probability model is established, while in the second this model is used to develop predictions, on the basis of which the required decision is made. The second of these stages was discussed in Section 2 and 3. We now discuss the first stage:

First observe that essentially there is no difference between snapshot data and time series data. In each case there is a sequence of observations of the variable. For snapshot data this ordering may represent the order in which the observations were made or simply the order in which they are listed. In any case, the assumption is made –and it is an assumption - that the measured variable is independent of this order. For time series data the ordering does represent the order in which the observations were made, and it is usually assumed that the variable is not independent of time.

There is however one fundamental difference. For snapshot data there is a population of real entities on which the measurements are made. If the data are collected on the whole population the probability distribution under random selection for the variable being

measured is directly observed and can be used as an empirical model, as described in section 2.4. More commonly a standard distribution of some sort is used as a model. In either case, the data are used to establish a probability model from which predictions can be made.

For time series data there is no such real population. For tomorrow's demand, there is no population of values from which one will be selected by chance. The **uncertainty** can however be **modelled**, using past data on which to base a probability model.

For snapshot data collected on a sample, the probability model is **inferred** from the sample results. This process is typically expressed in terms of 'describing the population', but this is true only to the extent that the chosen model adequately describes the population.

At the simplest level, the sample data are taken as the model. This is commonly done, for example, in newspaper reports of surveys. It is also done in textbooks in introductory probability using the 'frequentist' approach, particularly for examples using simple contingency tables. It is rarely made clear that the process of inference is involved.

For a numerical variable, an appropriate probability model type is chosen, then sample statistics used as estimates for the parameters of the model, and indeed **confidence intervals** for these parameters are obtained – based on the assumed model. It is worth repeating that the population is not itself described; the **assumed model** for the population is described. If that model type describes the population poorly, the analysis will be poor.

Introductory inference typically deals with means and proportions. There are good practical reasons for this: these are the parameters of the population in which the analyst is most commonly interested. And not surprisingly, they coincide with the 'best prediction' parameters.

For a nominal variable, since the best forecast is the mode, the probability of each outcome must be estimated. For a snapshot of a population under random selection, this probability is the population proportion for that outcome.

In order to forecast a numeric value, the population mean must be estimated, since this is the best forecast. It is also necessary to estimate the variability, as measured by the standard deviation, since this is used in computing the prediction interval.

The sample mean is the 'best estimate' for the population mean on the criteria of unbiasedness and minimum variance. These criteria correspond to those for 'best forecast'.

It is implicit in the above that a model must be assumed for the probability distribution of the variable. On the basis of this model, and assuming simple random sampling, an appropriate **sampling distribution** is derived and used to generate the estimates and the confidence intervals.

For a numeric variable it is usual to use the model

$$X_i \sim (m, s^2), \quad iid$$

Sample observations are made sequentially, so they form a time series, but as a consequence of the random selection the sequentiality does not matter: the time series is stationary. The model also assumes homoscedasticity. Both assumptions may be only partially true.

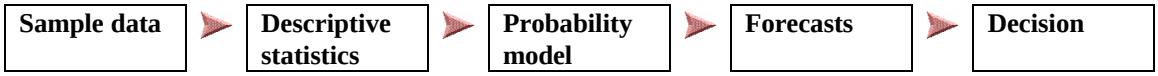
The central limit theorem tells us that a good model for the sampling distribution of the sample mean of a simple random sample from a large population (to give it its full name) is the standard normal, provided *n* is large enough; how large is 'large enough' depends on how appropriate the normal model is for *X* itself. In practice this theorem is always required, since the normal distribution is a model, so strictly *X* is **never** 'normally distributed'. Similarly, if σ is not known, but normality is a good model for *X*, then a *t* distribution is a good model for the standardised mean.

Using for example the normal model, a confidence interval for the population mean has limits

$$\bar{x} \pm z_{\alpha/2} \times s\sqrt{\frac{1}{n}}$$

5. Descriptive Statistics and Forecasting

The full process in a statistical analysis



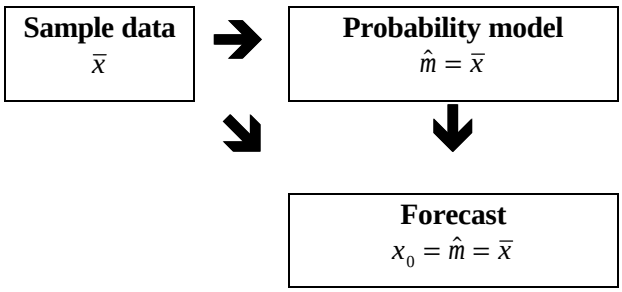
is to forecast an individual observation based on sample data, together with a specification of expected forecast quality. For a numeric variable, the aim is to provide a prediction interval for an individual observation. This will be 'best' achieved - under the criteria of zero absolute expected error and minimum expected squared error - if the sample mean is used to estimate the mean of the population model, μ, then this estimate used as the forecast.

If μ is estimated by some arbitrarily chosen number we can test this estimate on the sample, calculating the absolute mean error and mean squared error for the sample. Under this test, the sample mean performs best, giving zero absolute mean error and minimum mean squared error. That is, the sample mean gives the best forecast, under the same criteria as the mean of the population model. The sample mean is tested on the sample, the population mean is tested on the population.

To obtain a prediction interval for this forecast, the uncertainty in the estimate of the mean is combined with the variability assumed in the model for *X*. For the normal model with σ known, for example, a prediction interval has limits

$$\bar{x} \pm z_{\alpha/2} \times s\sqrt{\frac{1}{n}+1}$$

The calculation of the variance for the prediction interval combines the variance of the probability model and that of the sampling distribution by using Pythagoras. The process may be more appropriately shown rectilinearly:



Prediction intervals are sometimes introduced in the 'typical introductory text' under simple linear regression, when it is required to predict the value of *Y* for a given value of *x*; it is common to find both a confidence interval on the mean, and a prediction interval on the individual value. It is not pointed out that precisely the same approach can be applied in a univariate analysis.

In teaching regression there is usually some emphasis on prediction; it is recognised that the reason for carrying out a regression analysis is to predict the dependent variable. On the other hand, analysis of variance and cross tabulation are seen as extensions of basic statistics, in which subpopulations are compared in terms of a numeric or nominal variable respectively, and the predictive usage ignored. But the aim in each of these analyses is to identify a relationship which can provide better forecasts than are possible without it. Prediction is again the underlying purpose behind the analysis.

6. Concluding Remarks

It has been argued in this paper that the underlying purpose, often implicit rather than explicit, of every statistical analysis is to predict values of one or more variables, based on probability models for the variables. The models are in turn based on sample data. Using natural criteria, the 'best forecasts' for nominal and numeric variables are respectively the mode and mean. For a nominal variable, the quality of a prediction is measured by the probability of error; for a numeric variable, it is specified using a prediction interval.

The usefulness of a statistical analysis depends on the quality of the forecasts it leads to: if a statistical analysis leads to useful forecasts it is itself useful. An analysis which does not lead to useful predictions, however mathematically elegant, is of no practical use. On the other hand, if it shows that useful forecasts cannot be obtained, this result may also be of interest.

If it is accepted that this view of the underlying thrust of statistics is correct, then it is reasonable that texts should reflect this view. The predictive use of probability models, including the use of prediction intervals, should be emphasised. And, of utmost importance, the practical usefulness of results must be emphasised.

References

Foddy, WH (1988), *Elementary Applied Statistics for the Social Sciences*, Harper & Row, Sydney

Freeman, LC (1965), *Elementary Applied Statistics*

DeGroot, MH (1986), A Conversation With Persi Diaconis, *Statistical Science*, **1**, 3, 319-334

McLean, AL (1998), The Forecasting Voice: A Unified Approach to Teaching Statistics, *Proceedings of the Fifth International Conference on Teaching of Statistics, Singapore*, 1193-1199